**Income Qualification Course**

DESCRIPTION

Identify the level of income qualification needed for the families in Latin America. Problem Statement Scenario: Many social programs have a hard time ensuring that the right people are given enough aid. It’s tricky when a program focuses on the poorest segment of the population. This segment of the population can’t provide the necessary income and expense records to prove that they qualify. In Latin America, a popular method called Proxy Means Test (PMT) uses an algorithm to verify income qualification. With PMT, agencies use a model that considers a family’s observable household attributes like the material of their walls and ceiling or the assets found in their homes to classify them and predict their level of need. While this is an improvement, accuracy remains a problem as the region’s population grows and poverty declines. The Inter-American Development Bank (IDB)believes that new methods beyond traditional econometrics, based on a dataset of Costa Rican household characteristics, might help improve PMT’s performance.

**Following actions should be performed:**

• Identify the output variable.

• Understand the type of data.

• Check if there are any biases in your dataset.

• Check whether all members of the house have the same poverty level.

• Check if there is a house without a family head.

• Set poverty level of the members and the head of the house within a family.

• Count how many null values are existing in columns.

• Remove null value rows of the target variable.

• Predict the accuracy using random forest classifier.

• Check the accuracy using random forest with cross validation.

**APPROACH**

\* We can observe that we have a very small training data set as compared to test data set

\* Now let's find the target variable by referencing between the training and the test set

**Steps:**

**1. Data exploration:** First, we explore the dataset and try to understand the features and their distributions. Use summary statistics, histograms, other visualization techniques to get an idea of the data. We also identify missing values and outliers and decide on a strategy to handle them.

**2. Data preprocessing:** Next, we preprocess the data and prepare it for machine learning models. This includes handling missing values and outliers, encoding categorical features, and scaling numeric features. We also split the data into training and testing sets.

**3. Feature selection and engineering:** Once the data is preprocessed, we start selecting or engineering features that might be relevant for predicting income level. This involve transforming existing features, creating new features, or dropping irrelevant features.

**4. Model selection and training:** After feature selection and engineering, we start working on machine learning models. we use Random Forest Algorithm. Use cross-validation to evaluate the performance of each model and tune the hyperparameters.

**5. Model evaluation and interpretation:** Once we trained and optimized a model, we evaluated its performance on the testing set and checked the performance metrics of the models.

**Data Attributes Details**

1. ID = Unique ID

2. v2a1, Monthly rent payment

3. hacdor, =1 Overcrowding by bedrooms

4. rooms, number of all rooms in the house

5. hacapo, =1 Overcrowding by rooms

6. v14a, =1 has bathroom in the household

7. refrig, =1 if the household has a refrigerator

8. v18q, owns a tablet

9. v18q1, number of tablets household owns

10. r4h1, Males younger than 12 years of age

11. r4h2, Males 12 years of age and older

12. r4h3, Total males in the household

13. r4m1, Females younger than 12 years of age

14. r4m2, Females 12 years of age and older

15. r4m3, Total females in the household

16. r4t1, persons younger than 12 years of age

17. r4t2, persons 12 years of age and older

18. r4t3, Total persons in the household

19. tamhog, size of the household

20. tamviv, number of persons living in the household

21. escolari, years of schooling

22. rez\_esc, Years behind in school

23. hhsize, household size

24. paredblolad, =1 if predominant material on the outside wall is block or brick

25. paredzocalo, "=1 if predominant material on the outside wall is socket (wood, zinc or

absbesto"

26. paredpreb, =1 if predominant material on the outside wall is prefabricated or cement

27. pareddes, =1 if predominant material on the outside wall is waste material

28. paredmad, =1 if predominant material on the outside wall is wood

29. paredzinc, =1 if predominant material on the outside wall is zink

30. paredfibras, =1 if predominant material on the outside wall is natural fibers

31. paredother, =1 if predominant material on the outside wall is other

32. pisomoscer, "=1 if predominant material on the floor is mosaic, ceramic, terrazo"

33. pisocemento, =1 if predominant material on the floor is cement

34. pisoother, =1 if predominant material on the floor is other

35. pisonatur, =1 if predominant material on the floor is natural material

36. pisonotiene, =1 if no floor at the household

37. pisomadera, =1 if predominant material on the floor is wood

38. techozinc, =1 if predominant material on the roof is metal foil or zink

39. techoentrepiso, "=1 if predominant material on the roof is fiber cement, mezzanine "

40. techocane, =1 if predominant material on the roof is natural fibers

41. techootro, =1 if predominant material on the roof is other

42. cielorazo, =1 if the house has ceiling

43. abastaguadentro, =1 if water provision inside the dwelling

44. abastaguafuera, =1 if water provision outside the dwelling

45. abastaguano, =1 if no water provision

46. public, "=1 electricity from CNFL, ICE, ESPH/JASEC"

47. planpri, =1 electricity from private plant

48. noelec, =1 no electricity in the dwelling

49. coopele, =1 electricity from cooperative

50. sanitario1, =1 no toilet in the dwelling

51. sanitario2, =1 toilet connected to sewer or cesspool

52. sanitario3, =1 toilet connected to septic tank

53. sanitario5, =1 toilet connected to black hole or letrine

54. sanitario6, =1 toilet connected to other system

55. energcocinar1, =1 no main source of energy used for cooking (no kitchen)

56. energcocinar2, =1 main source of energy used for cooking electricity

57. energcocinar3, =1 main source of energy used for cooking gas

58. energcocinar4, =1 main source of energy used for cooking wood charcoal

59. elimbasu1, =1 if rubbish disposal mainly by tanker truck

60. elimbasu2, =1 if rubbish disposal mainly by botan hollow or buried

61. elimbasu3, =1 if rubbish disposal mainly by burning

62. elimbasu4, =1 if rubbish disposal mainly by throwing in an unoccupied space

63. elimbasu5, "=1 if rubbish disposal mainly by throwing in river, creek or sea"

64. elimbasu6, =1 if rubbish disposal mainly other

65. epared1, =1 if walls are bad

66. epared2, =1 if walls are regular

67. epared3, =1 if walls are good

68. etecho1, =1 if roof are bad

69. etecho2, =1 if roof are regular

70. etecho3, =1 if roof are good

71. eviv1, =1 if floor are bad

72. eviv2, =1 if floor are regular

73. eviv3, =1 if floor are good

74. dis, =1 if disable person

75. male, =1 if male

76. female, =1 if female

77. estadocivil1, =1 if less than 10 years old

78. estadocivil2, =1 if free or coupled uunion

79. estadocivil3, =1 if married

80. estadocivil4, =1 if divorced

81. estadocivil5, =1 if separated

82. estadocivil6, =1 if widow/er

83. estadocivil7, =1 if single

84. parentesco1, =1 if household head

85. parentesco2, =1 if spouse/partner

86. parentesco3, =1 if son/doughter

87. parentesco4, =1 if stepson/doughter

88. parentesco5, =1 if son/doughter in law

89. parentesco6, =1 if grandson/doughter

90. parentesco7, =1 if mother/father

91. parentesco8, =1 if father/mother in law

92. parentesco9, =1 if brother/sister

93. parentesco10, =1 if brother/sister in law

94. parentesco11, =1 if other family member

95. parentesco12, =1 if other non family member

96. idhogar, Household level identifier

97. hogar\_nin, Number of children 0 to 19 in household

98. hogar\_adul, Number of adults in household

99. hogar\_mayor, # of individuals 65+ in the household

100. hogar\_total, # of total individuals in the household

101. dependency, Dependency rate, calculated = (number of members of the household

younger than 19 or older than 64)/(number of member of household between 19 and 64)

102. edjefe, years of education of male head of household, based on the interaction of

escolari (years of education), head of household and gender, yes=1 and no=0

103. edjefa, years of education of female head of household, based on the interaction of

escolari (years of education), head of household and gender, yes=1 and no=0

104. meaneduc,average years of education for adults (18+)

105. instlevel1, =1 no level of education

106. instlevel2, =1 incomplete primary

107. instlevel3, =1 complete primary

108. instlevel4, =1 incomplete academic secondary level

109. instlevel5, =1 complete academic secondary level

110. instlevel6, =1 incomplete technical secondary level

111. instlevel7, =1 complete technical secondary level

112. instlevel8, =1 undergraduate and higher education

113. instlevel9, =1 postgraduate higher education

114. bedrooms, number of bedrooms

115. overcrowding, # persons per room

116. tipovivi1, =1 own and fully paid house

117. tipovivi2, "=1 own, paying in installments"

118. tipovivi3, =1 rented

119. tipovivi4, =1 precarious

120. tipovivi5, "=1 other(assigned, borrowed)"

121. computer, =1 if the household has notebook or desktop computer

122. television, =1 if the household has TV

123. mobilephone, =1 if mobile phone

124. qmobilephone, # of mobile phones

125. lugar1, =1 region Central

126. lugar2, =1 region Chorotega

127. lugar3, =1 region PacÃ fico central

128. lugar4, =1 region Brunca

129. lugar5, =1 region Huetar AtlÃ¡ntica

130. lugar6, =1 region Huetar Norte

131. area1, =1 zona urbana

132. area2, =2 zona rural

133. age= Age in years

134. SQBescolari= escolari squared

135. SQBage, age squared

136. SQBhogar\_total, hogar\_total squared

137. SQBedjefe, edjefe squared

138. SQBhogar\_nin, hogar\_nin squared

139. SQBovercrowding, overcrowding squared

140. SQBdependency, dependency squared

141. SQBmeaned, square of the mean years of education of adults (>=18) in the

household

142. agesq= Age squared

**LIST OF FEATURES USED**

1. r4h1, Males younger than 12 years of age
2. r4m1, Females younger than 12 years of age
3. r4t1, persons younger than 12 years of age
4. escolari, years of schooling
5. hogar\_nin, Number of children 0 to 19 in household
6. rez\_esc, Years behind in school
7. dependency, Dependency rate, calculated = (number of members of the household younger than 19 or older than 64)/(number of member of household between 19 and 64)
8. edjefe, years of education of male head of household, based on the interaction of escolari (years of education), head of household and gender, yes=1 and no=0
9. edjefa, years of education of female head of household, based on the interaction of escolari (years of education), head of household and gender, yes=1 and no=0
10. hacdor, =1 Overcrowding by bedrooms
11. hacapo, =1 Overcrowding by rooms
12. pareddes, =1 if predominant material on the outside wall is waste material
13. paredfibras, =1 if predominant material on the outside wall is natural fibers
14. pisonatur, =1 if predominant material on the floor is natural material
15. pisonotiene, =1 if no floor at the household
16. techocane, =1 if predominant material on the roof is natural fibers
17. abastaguano, =1 if no water provision
18. noelec, =1 no electricity in the dwelling
19. sanitario1, =1 no toilet in the dwelling
20. sanitario5, =1 toilet connected to black hole or letrine
21. energcocinar1, =1 no main source of energy used for cooking (no kitchen)
22. age= Age in years
23. v2a1, Monthly rent payment
24. meaneduc, average years of education for adults (18+)
25. Target

**EDA**

1. **Remove squared variables**

In the dataset, we remove squared variables (i.e., variables that are squared or raised to a power greater than 1) from the analysis for several reasons:

Interpretability

Multicollinearity

Overfitting

1. classifies the columns (variables) into four categories based on the number of unique values (nunique) they contain:

**garbage\_col**: These are columns with only one unique value, which provide no useful information for the analysis. They are considered garbage columns and will be removed from the analysis.

**yes\_no\_columns**:['hacdor','hacapo','v14a','refrig','v18q','r4h2'] These are columns with only two unique values, which are often binary variables (e.g., 0/1, True/False). These variables are typically treated as categorical variables with two categories and are often useful in modeling.

**nominal\_cols**:['rooms','r4h1','r4h3','r4m1','r4m2','r4m3','r4t1','r4t2'] These are columns with more than two unique values, but less than or equal to 38 unique values. These columns are considered nominal categorical variables, which have no intrinsic ordering or hierarchy among their categories. These variables are often encoded using one-hot encoding for modeling.

**continous**\_cols:['age', 'v2a1'] These are columns with more than 38 unique values, and are considered continuous variables. These variables can take on any value within a range, and are often used in regression and other modeling techniques that assume a continuous distribution.

1. We perform univariate and bivariate analysis of continuous variables with target variables.

This analysis provides a quick overview of the distribution and range of the continuous variables in the dataset. It can help identify outliers, missing values, or potential data quality issues that may need further investigation before modeling.

1. We performs the chi-square test for independence between a categorical variable (columns) and the target variable ('Target') in a given DataFrame (df), From chi-square test we select the variable which is significant(i.e, 'r4h1', 'r4m1', 'r4t1', 'escolari', 'hogar\_nin', 'rez\_esc', 'dependency', 'edjefe', 'edjefa')
2. We used pivot table to analyze the relationship between the poverty level and slect\_col\_nominals i.e significant variable. The table shows the count of households, categorized by the significant variable ('r4h1', 'r4m1', 'r4t1', 'escolari', 'hogar\_nin', 'rez\_esc', 'dependency', 'edjefe', 'edjefa'), for each target category. The target categories are represented by the values 1, 2, 3, and 4.
3. Ater that we used the chi-square test for independence between yes\_no\_columns(columns) and the target variable ('Target') in a given DataFrame (df). From chi-square test we select the variable which is significant(i.e, "hacdor" ,"hacapo" ,"pareddes" ,"paredfibras" ,"pisonatur" ,"pisonotiene" ,"techocane" ,"abastaguano" ,"noelec" ,"sanitario1" ,"sanitario5" ,"energcocinar1" )

**ALGORITHM USED**

\* x\_features, y\_features: The first parameter is the dataset you're selecting to use.

\* train\_size: This parameter sets the size of the training dataset. There are three options: None, which is the default, Int, which requires the exact number of samples, and float, which ranges from 0.1 to 1.0.

\* test\_size: This parameter specifies the size of the testing dataset. The default state suits the training size. It will be set to 0.25 if the training size is set to default.

\* random\_state: The default mode performs a random split using np.random. Alternatively, you can add an integer using an exact number.

* **Random Forest Algorithm**

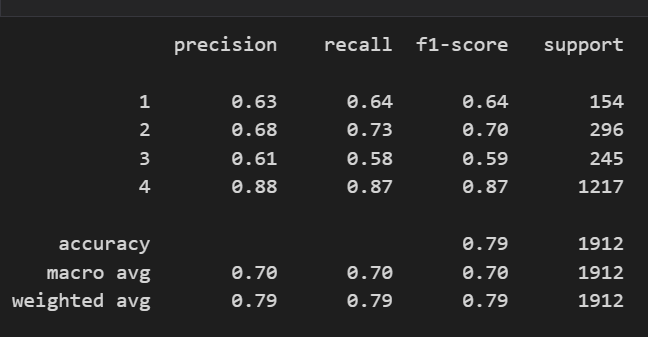
Random Forest Classifier is a type of ensemble learning algorithm that combines multiple decision trees to form a more robust and accurate model. It works by training multiple decision trees on different subsets of the data and then averaging their predictions. This method reduces overfitting and increases the generalization power of the model.

After that we tuned the hyperparameters of the Random Forest Classifier to optimize the performance of the model.

**SMOTE**

We use Random Forest Classifier with imbalanced datasets, so to balance the class distribution we use SMOTE. This helps prevent the model from being biased towards the majority class and improve its accuracy on the minority class.

Evaluation Metrix



**Prediction Result**

**Model gives accuracy 79% but as our data is imbalanced along with accuracy we need to check precision, recall and f1-score to get more accurate predictions.**

**F1 – Score = 64%**